Machine Learning

Haoyu Zhen

March 4, 2022

Contents

1 Foundations **2**

1.1 Model evaluation: 2

1.2 Performance 2

1.3 Bias-Variance Decomposition 2

2 Regression **2**

2.1 Linear Regression 2

2.2 Ridge Regression 3

2.3 Lasso Regression 4

2.4 Logistic Regression 4

Acknowledgement

These Notes contain material developed and copyright by:

- *AI-2611, Machine Learning*, © 2022 by Bingbing Ni, Shanghai Jiao Tong University.
- *Understanding Machine Learning*, © 2014 by Shai Shalev-Shwartz and Shai Ben-David.
- *The Elements of Statistical Learning*, © 2017 by Trevor Hastie, Robert Tibshirani and Jerome Friedman.

1 Foundations

1.1 Model evaluation:

Hold-out, cross validation and bootstrap.
For cross validation, we often let the numbers of the folds be 10. And in bootstrap, the equation $\lim_{n\to\infty}(1 - 1/m)^m = 1/e$ is used to analyse the probality.

1.2 Performance

Now we consider that:

	prediction+	prediction-
Actual	1	0
1	TP	FP
0	FN	TN

Definition 1.1 (Sensitivity and FPR)**.**

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{TN + FP}.$$

Then we introduce **ROC** space and **AUC**.

Definition 1.2 (Precision and recall)**.**

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN}.$$

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{\beta^2 \times P + R}.$$

β depends on the preference of Precision and Recll.

1.3 Bias-Variance Decomposition

$$\begin{aligned} E(f; D) &= bias^2(x) + var(x) + \varepsilon^2 \\ &= (\bar{f}(x) - y)^2 + \mathbb{E}_D[f(x; D) - \bar{f}(x)] + \mathbb{E}_D[(y_D - y)^2] \end{aligned}$$

2 Regression

2.1 Linear Regression

The hypothesis class of linear regression predictors is simply the set of linear functions,

$$\mathcal{H}_{reg} = \{\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b : \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

Intuitively,

$$\mathcal{L}_{\mathcal{S}}(h) = \frac{1}{m}\sum_{i=1}^{m}(h(\boldsymbol{x}) - \boldsymbol{y})^2, \forall h \in \mathcal{H}_{reg}.$$

To minimize the loss function, we need to solve $A\boldsymbol{w} = \boldsymbol{b}$ where $A \stackrel{def}{=\!=} \sum \boldsymbol{x}_i\boldsymbol{x}_i^T = XX^T$ and $\boldsymbol{b} \stackrel{def}{=\!=} \sum y_i\boldsymbol{x}_i = X^T\boldsymbol{y}$. If A is invertible then the solution is $w = A^{-1}\boldsymbol{b}$.

Theorem 2.1.

$$\omega = (X^TX)^{-1}X^T\boldsymbol{y}.$$

If the training instances do not span the entire space of $\mathbb{R}^d$ then A is not invertible.

Theorem 2.2. Using A's eigenvalue decomposition,we could write A as VD^+V^T where D is a diagnonal matrix and V is an orthonormal matrix. Define D^+ to be the diagonal matrix such that $D^+_{i,i} = 0$ if $D_{i,i} = 0$ otherwise $D^+_{i,i} = 1/D_{i,i}$. Then,

$$A\hat{\boldsymbol{w}} = \boldsymbol{b}$$

where $\hat{\boldsymbol{w}} = VD^+V^T\boldsymbol{b}$

Proof.

$$A\hat{\omega} = AA^+\boldsymbol{b} = VDV^TVD^+V^T\boldsymbol{b} = VDD^+V^T\boldsymbol{b} = \sum_{i:D_{i,i}\neq 0}\boldsymbol{v}_i\boldsymbol{v}_i^T\boldsymbol{b}.$$

That is, $A\hat{\omega}$ is the projection of b onto the span of those vectors v_i for which $D_{i,i} \neq 0$. Since the linear span of $x_1, \cdots, x_m$ is the same as the linear span of those v_i, and b is in the linear span of the x_i, we obtain that $A\hat{w} = b$, which concludes our argument. □

Remark 2.1. Indeed we always use the **Gradient Descent** method to optimize the loss function.

Linear regression for polynomial regression tasks $\mathcal{H}^n_{poly} = \{x \mapsto p(x)\}$ where $\psi(x) = (1, x, x^2, \cdots, x^n)$ and $p(\psi(x)) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$.

2.2 Ridge Regression

To ameliorate the effect of the invertible matrix, we could introduce the regularization.

Definition 2.1 (Regularized Loss).

$$R(w) = \lambda\|w\|^2.$$

Now the loss function reads:

$$\mathcal{L} = \mathcal{L}_{\mathcal{S}}(w) + R(w) = \frac{1}{m}\sum_{i=1}^{m}(h(\boldsymbol{x}) - \boldsymbol{y})^2 + \lambda\|w\|^2.$$

2.3 Lasso Regression

2.4 Logistic Regression